(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 104281694 A
(43)申请公布日 2015.01.14

(21)申请号 201410537881.0

(22)申请日 2014.10.13

(71)申请人 安徽华贞信息科技有限公司
地址 230000 安徽省合肥市高新区黄山路602号国家大学科技园A502

(72)发明人 贾岩

(74)专利代理机构 合肥市长远专利代理事务所（普通合伙） 34119
代理人 程笃庆 黄乐瑜

(51)Int.Cl.
G06F 17/30(2006.01)
G06F 17/27(2006.01)

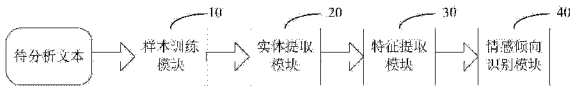
权利要求书1页 说明书2页 附图2页

(54)发明名称

一种文本情感倾向分析系统

(57)摘要

本发明公开了一种文本情感倾向分析系统，包括：样本训练模块，用于接收待分析文本，训练样本，获取判别模板；实体提取模块，对待判别文本实体提取，过滤不含实体的文本；特征提取模块，提取文本中的倾向性相关特征；情感倾向识别模块，利用最大熵方法判别文本倾向性。本发明采集用户企业所在领域的论坛、博客，提取网页中的文本，通过文本情感倾向分析获得文本的情感倾向以及针对的实体，并自动生成企业和竞争对手形象变化图表，以提高利用文本分类来判断文本情感倾向性的准确率。



CN 104281694 A

1. 一种文本情感倾向分析系统，其特征在于，包括：

样本训练模块，用于接收待分析文本，训练样本，获取判别模板；

实体提取模块，对待判别文本实体提取，过滤不含实体的文本；

特征提取模块，提取文本中的倾向性相关特征；

情感倾向识别模块，利用最大熵方法判别文本倾向性。

2. 根据权利要求 1 所述的文本情感倾向分析系统，其特征在于，所述倾向性相关特征包括：极性词、维度词、修饰词、否定词。

3. 根据权利要求 1 所述的文本情感倾向分析系统，其特征在于，对文本进行倾向分析的之前建立实体词典、极性词典、维度词典、修饰词词典以及其它相关词典。

4. 根据权利要求 1 所述的文本情感倾向分析系统，其特征在于，所述实体提取模块，具体用于：

预处理；

项权重的计算；

根据预处理的训练集；

学习建模，构建出分类器；

利用测试集文档按一定的测试方法测试建立好的分类器的性能，并不断反馈、学习提高该分类器性能，直至达到预定目标。

5. 根据权利要求 4 所述的文本情感倾向分析系统，其特征在于，所述预处理具体为：根据采用的分类模型将文档集表示成易于计算机处理的形式。

6. 根据权利要求 4 所述的文本情感倾向分析系统，其特征在于，所述项权重的计算，具体为：根据适宜的权重计算方法表示文档中各项的重要性。

7. 根据权利要求 1 所述的文本情感倾向分析系统，其特征在于，所述特征提取模块，具体用于：

通过关键词抽取或者特征提取文本中的特征词；

通过向量空间模型将文档向量化；

计算文档之间的相似度，并选择合适算法进行聚类。

一种文本情感倾向分析系统

技术领域

[0001] 本发明涉及数据网络技术领域，尤其涉及一种文本情感倾向分析系统。

背景技术

[0002] 为了搜索和竞争情报系统分析企业和产品形象，通常会使用倾向性分析，而按照倾向性的程度将文本分成几类。由于文本的倾向性不仅由极性词、程度词等这些词来决定，还和这些词的相对位置以及和实体词的关系有关，而文本分类只能考虑词的特征，所以目前利用文本分类来判断文本情感倾向性的一些方法准确率都较低。

发明内容

[0003] 为了解决背景技术中存在的技术问题，本发明提出了一种文本情感倾向分析系统，提高利用文本分类来判断文本情感倾向性的准确率。

[0004] 本发明提出的一种文本情感倾向分析系统，包括：

[0005] 样本训练模块，用于接收待分析文本，训练样本，获取判别模板；

[0006] 实体提取模块，对待判别文本实体提取，过滤不含实体的文本；

[0007] 特征提取模块，提取文本中的倾向性相关特征；

[0008] 情感倾向识别模块，利用最大熵方法判别文本倾向性。

[0009] 优选地，所述倾向性相关特征包括：极性词、维度词、修饰词、否定词。

[0010] 优选地，对文本进行倾向分析的之前建立实体词典、极性词典、维度词典、修饰词词典以及其它相关词典。

[0011] 优选地，所述实体提取模块，具体用于：

[0012] 预处理；

[0013] 项权重的计算；

[0014] 根据预处理的训练集；

[0015] 学习建模，构建出分类器；

[0016] 利用测试集文档按一定的测试方法测试建立好的分类器的性能，并不断反馈、学习提高该分类器性能，直至达到预定目标。

[0017] 优选地，所述预处理具体为：根据采用的分类模型将文档集表示成易于计算机处理的形式。

[0018] 优选地，所述项权重的计算，具体为：根据适宜的权重计算方法表示文档中各项的重要性。

[0019] 优选地，所述特征提取模块，具体用于：

[0020] 通过关键词抽取或者特征提取文本中的特征词；

[0021] 通过向量空间模型将文档向量化；

[0022] 计算文档之间的相似度，并选择合适算法进行聚类。

[0023] 本发明中，采集用户企业所在领域的论坛、博客，提取网页中的文本，通过文本情

感倾向分析获得文本的情感倾向以及针对的实体（企业、企业产品、竞争对手等），并自动生成企业和竞争对手形象变化图表，以提高利用文本分类来判断文本情感倾向性的准确率。

附图说明

[0024] 图 1 为本发明实施例提出的一种文本情感倾向分析系统；

[0025] 图 2 为图 1 中文本分类模块的功能图；

[0026] 图 3 为图 1 中文本聚类模块的功能图。

具体实施方式

[0027] 如图 1 所示，本发明实施例提出了一种文本情感倾向分析系统，包括：样本训练模块 10，用于接收待分析文本，训练样本，获取判别模板；实体提取模块 20，对待判别文本实体提取，过滤不含实体的文本；特征提取模块 30，提取文本中的倾向性相关特征（极性词、维度词、修饰词、否定词等）；情感倾向识别模块 40，利用最大熵方法判别文本倾向性。另外，对文本进行倾向分析的之前建立实体词典、极性词典、维度词典、修饰词词典以及其它相关词典。

[0028] 其中，实体提取模块 20 的功能如图 2 所示，包括：首先是预处理，根据采用的分类模型将文档集表示成易于计算机处理的形式；其次是项权重的计算，根据适宜的权重计算方法表示文档中各项的重要性；再次是根据预处理的训练集（已预知类别的文档）学习建模，构建出分类器；最后利用测试集文档按一定的测试方法测试建立好的分类器的性能，并不断反馈、学习提高该分类器性能，直至达到预定目标。

[0029] 其中，特征提取模块 30 的功能如图 3 所示，包括：通过关键词抽取或者特征提取文本中的特征词，然后通过向量空间模型将文档向量化，最后计算文档之间的相似度，并选择合适算法进行聚类。

[0030] 以上所述，仅为本发明较佳的具体实施方式，但本发明的保护范围并不局限于此，任何熟悉本技术领域的技术人员在本发明揭露的技术范围内，根据本发明的技术方案及其发明构思加以等同替换或改变，都应涵盖在本发明的保护范围之内。

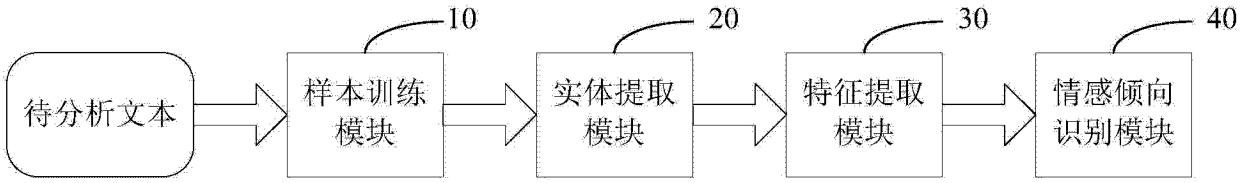

待分析文本
10
样本训练模块
20
实体提取模块
30
特征提取模块
40
情感倾向识别模块

图 1

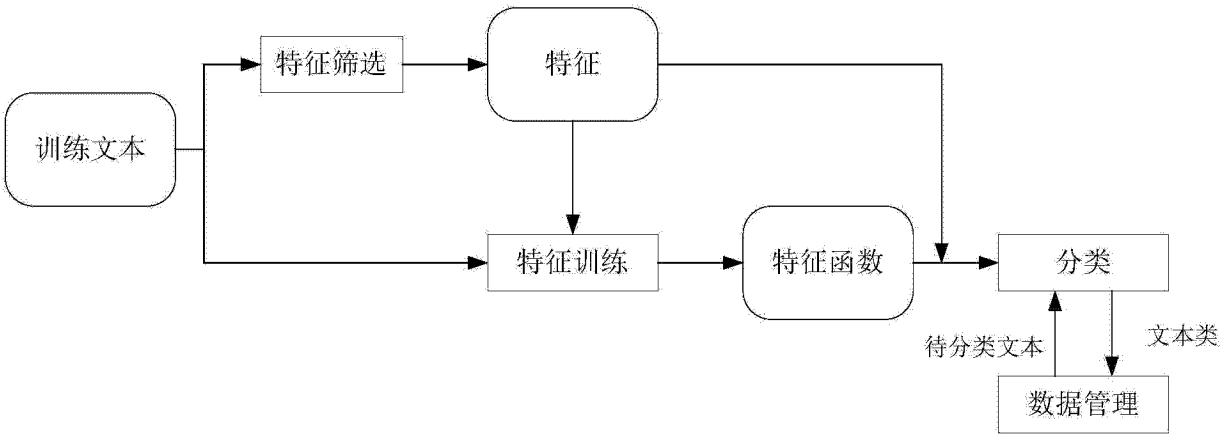

训练文本
特征筛选
特征
特征训练
特征函数
分类
待分类文本
文本类
数据管理

图 2

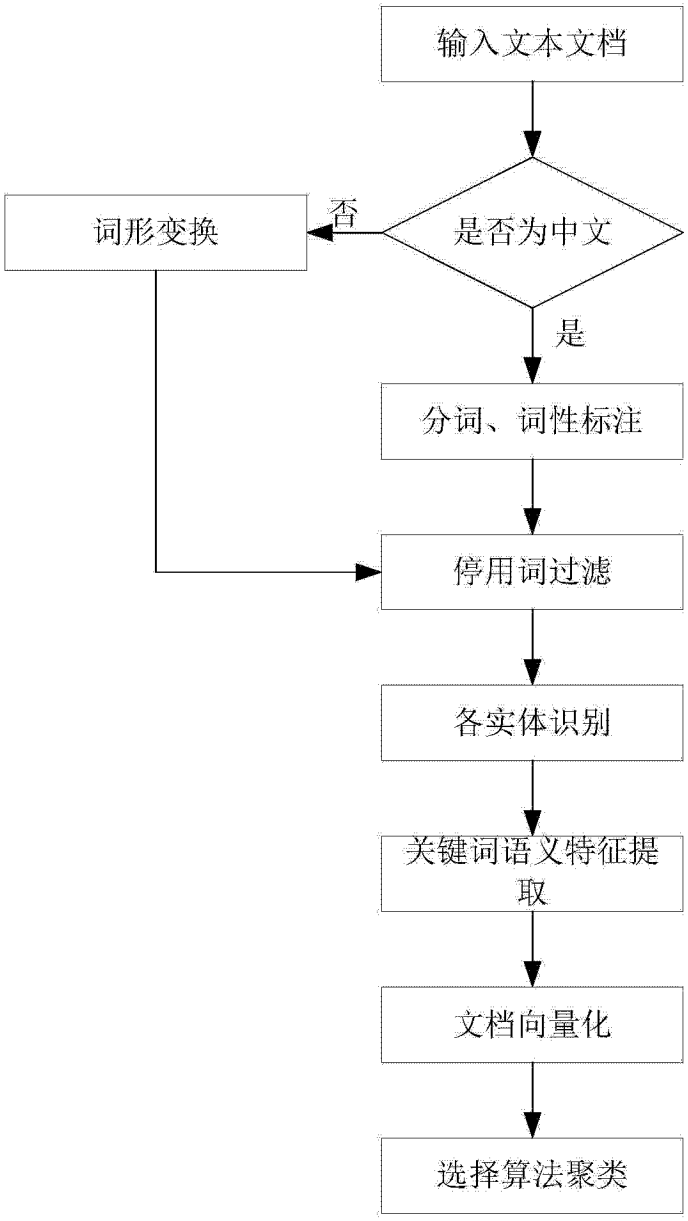

输入文本文档
是否为中文
否
词形变换
是
分词、词性标注
停用词过滤
各实体识别
关键词语义特征提取
文档向量化
选择算法聚类

图 3